

Robust Distributed Averaging: When are Potential-Theoretic Strategies Optimal?

Ali Khanafer and Tamer Başar

Abstract

We study the interaction between a network designer and an adversary over a dynamical network. The network consists of nodes performing continuous-time distributed averaging. The adversary strategically disconnects a set of links to prevent the nodes from reaching consensus. Meanwhile, the network designer assists the nodes in reaching consensus by changing the weights of a limited number of links in the network. We formulate two Stackelberg games to describe this competition where the order in which the players act is reversed in the two problems. Although the canonical equations provided by the Pontryagin's maximum principle seem to be intractable, we provide an alternative characterization for the optimal strategies that makes connection to potential theory. Finally, we provide a sufficient condition for the existence of a saddle-point equilibrium for the underlying zero-sum game.

I. INTRODUCTION

Various physical and biological phenomena where global patterns of behavior stem from local interactions have been modeled using linear distributed averaging dynamics. In such dynamics an agent updates its value as a linear combination of the values of its neighbors. Averaging dynamics is the basic building block in many multi-agent systems, and it is widely used whenever an application requires multiple agents, who are graphically constrained, to synchronize their measurements. Examples include formation control, coverage, distributed estimation and

A preliminary version of this work will was presented at the IEEE Conference on Decision and Control (CDC), Florence, Italy 2013 [1]. This work was supported in part by an AFOSR MURI Grant FA9550-10-1-0573.

Ali Khanafer and Tamer Başar are with the Coordinated Science Laboratory, ECE Department, University of Illinois at Urbana-Champaign, USA [khanafe2](mailto:khanafe2@illinois.edu), basar1@illinois.edu

optimization, and flocking [2]–[4]. Besides engineering, linear distributed averaging finds applications in other fields as well. For instance, social scientists use averaging to describe the evolution of opinions in networks [5].

In practice, communication among agents is prone to different types of non-idealities which can affect the convergence properties of the associated distributed algorithms. Transmission delays [6], noisy links [7], [8], and quantization [9] are some examples of non-idealities that are due to the physical nature of the application. In addition to physical restrictions, researchers have also studied averaging dynamics in the presence of malicious nodes in the network [10], [11].

In [12], we explored the effect of an external adversary who attempts to prevent the nodes from reaching consensus by launching network-wide attacks. When the adversary is capable of disconnecting certain links in the network, we derived the optimal strategy of the adversary and demonstrated that it admits a potential-theoretic analogy. In this paper, we also introduce a network designer that attempts to counter the effect of the adversary and help the nodes reach consensus. The designer is capable of changing the weights of certain links. Both the adversary and the designer are constrained by their physical capabilities, e.g., battery life and communication range. To capture such constraints, we allow the adversary and the designer to affect only a fixed number of links at any point in time. The conflicting objectives of the designer and the adversary calls for a game-theoretic formulation to study their interaction.

Such a competition between a network designer and an adversary can occur in various practical applications. For example, in a wireless network, the link weights in such a network represent the capacities of the corresponding links. The designer can modify the capacity of a certain link using various communication techniques such as introducing parallel channels between two nodes as in orthogonal frequency division multiple access (OFDMA) networks [13]. In OFDMA networks, the number of parallel links between two nodes is usually limited [14]. To capture this limitation, we limit the amount by which the designer can increase the capacity of a given link. The adversary can be a jammer who is capable of breaking links by injecting high noise signals that disrupt the communication among nodes. The adversary is assumed to have sufficient transmit power to disrupt the communication over any link, no matter what the number of parallel channels is.

Our model in this paper is different from the models in the current literature in two ways: (i) the adversary and the designer compete over a dynamical network. This is different from

the problems studied in the computer science and economics communities where the network is usually static [15]; (ii) the players in our model are constrained and do not have an infinite budget. This enables us to model practical scenarios more closely rather than allowing the malicious behavior to be unrestricted as in [10], [16], [17], where it is assumed that the network contains nodes that are misbehaving. In addition, those papers focus on finding necessary and sufficient conditions for the network to reach consensus in the presence of malicious nodes, and observability theory is the main tool used to study such problems. Here, we assume that all the nodes are normal, and we focus on identifying the links that are of importance to the adversary and the designer who have conflicting objectives. This requires us to borrow tools from differential games and optimal control theory.

The main goal of this work is to derive optimal strategies for the designer and the adversary. By modeling the adversary as a strategic player and deriving optimal defense strategies, we guarantee robustness against worst-case attacks, unlike existing approaches in which attacks on links were modeled as random failures [18]. Because the order in which the players act affects the resulting utilities, we formulate two Stackelberg games based on the order of play, allowing a different player to have the *first-move-advantage*. When the adversary is allowed to play first, he is capable of restricting the available actions of the designer since some links will disappear from the network. Hence, if we were to cast the problem as a zero-sum game between the players, we should not expect the existence of a saddle-point equilibrium (SPE) in pure strategies. The question we would like to answer is then: *are there scenarios where the order of play does not affect the eventually attained utilities of the players, which leads to the existence of an SPE?*

Accordingly, the contributions of this paper are as follows. We capture the interaction between the designer and the adversary by formulating two separate problems. In the min–max problem, the designer declares a strategy first to which the adversary reacts by its optimal response. The second problem is a max–min one where the order of play is reversed. Assuming the controllers do not switch infinitely many times over a finite interval among the available actions, we derive the optimal strategies for both problems in terms of potential-theoretic quantities by working directly with the utility functional. Furthermore, we demonstrate that the derived strategies satisfy the necessary conditions provided by the maximum principle (MP). Finally, we derive a sufficient condition guaranteeing the existence of an SPE.

The rest of the paper is organized as follows. In Section II we describe the min–max and

max–min problems. In Section III, we derive the Stackelberg strategies and show that they satisfy the MP. We provide a sufficient condition for the existence of an SPE in Section IV. We end the paper with concluding remarks and delineation of future research directions of Section V. An Appendix includes a proof of one of the theorems and two technical results.

Notation and Terminology

We will use $\sum_{j>i}(\cdot)$ to mean $\sum_{j=2}^n \sum_{i=1}^{j-1}(\cdot)$, $[\cdot]^T$ to denote the transpose of a vector or a matrix $[\cdot]$, and $\mathbf{1}$ to denote the n -dimensional column vector of 1's. The Euclidean norm of a vector is denoted by $\|\cdot\|_2$ and the ℓ_1 -norm of a vector is denoted by $\|\cdot\|_1$. The absolute value of a scalar variable is denoted by $|\cdot|$, which we also use to denote the cardinality of a set—the intended use of this operator will be clear from the context. The (i, j) -th element of a matrix X is denoted by X_{ij} . We will often use x to refer to a function or its value at a given time instant; the context should make the distinction clear. We will use the words “strategy” and “action” interchangeably; since we are seeking optimal open-loop strategies in this paper, both terms are equivalent. A graph is a pair $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where \mathcal{N} is the set of nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. An edge from node $i \in \mathcal{N}$ to node $j \in \mathcal{N}$ is denoted by e_{ij} , i.e., $e_{ij} := (i, j)$. A graph is called undirected if $e_{ij} \in \mathcal{E}$ if and only if $e_{ji} \in \mathcal{E}$. A path is a collection of nodes $\{i_1, \dots, i_l\} \subseteq \mathcal{N}$, $l \in \mathbb{Z}_{>1}$, such that $e_{i_k i_{k+1}} \in \mathcal{E}$ for all $k \in \{1, \dots, l-1\}$. We call an undirected graph *connected* if it contains a path between any two nodes in \mathcal{N} . Given an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we define the projection operator $\Phi : \mathcal{E} \times \mathbb{R} \rightarrow \mathcal{E}$ such that $\Phi((e, r)) = e$, for some $(e, r) \in \mathcal{E} \times \mathbb{R}$. When applied to a set $S \subset \mathcal{E} \times \mathbb{R}$, the mapping Φ is defined as follows:

$$\Phi(S) = \begin{cases} \bigcup_{(e,r) \in S} \Phi((e, r)), & S \neq \emptyset \\ 0, & S = \emptyset \end{cases}$$

Given $S \subset \mathcal{E} \times \mathbb{R}$, with $|S| = k$, let $\pi(S) = \{(e_1, r_1), \dots, (e_k, r_k)\}$, where $r_i \in \mathbb{R}$ and $e_i \in \mathcal{E}$ for all $i \in \{1, \dots, k\}$, be an ordering of the elements of S such that $r_1 \leq \dots \leq r_k$. Then, given $\ell \in \mathbb{Z}_{\geq 0}$, we define the set operator $\Phi_\ell : \mathcal{E} \times \mathbb{R} \rightarrow \mathcal{E}$ as:

$$\Phi_\ell(S) = \begin{cases} \Phi(S), & \ell > k \\ \{e_1, \dots, e_\ell\}, & 0 < \ell \leq k \\ 0, & \ell = 0 \text{ or } k = 0 \end{cases}$$

Throughout the paper, we will be dealing with undirected graphs. Although both e_{ij}, e_{ji} belong to the set of edges \mathcal{E} in such graphs, we do not distinguish between the two edges, and we treat them as a single edge. As a result, in any set defined over $\mathcal{E} \times \mathbb{R}$, we include a *single* tuple (e_{ij}, r_{ij}) , $r_{ij} \in \mathbb{R}$, to represent both edges.

II. PROBLEM FORMULATION

Consider a connected network of n nodes and m links described by a weighted undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. The value, or state, of the nodes at time instant $t \in \mathbb{R}_{\geq 0}$ is given by $x(t) = [x_1(t), \dots, x_n(t)]^T$. The nodes start with an initial value $x(0) = x_0$, and they are interested in computing the average of their initial values, $x_{\text{avg}} = \frac{1}{n} \sum_{i=1}^n x_i(0)$, via local averaging. We consider the continuous-time averaging dynamics given by

$$\dot{x}(t) = Ax(t), \quad x(0) = x_0, \quad (1)$$

where the matrix A , $A_{ij} = a_{ij} \in \mathbb{R}$, has the following properties:

$$\begin{aligned} A &= A^T, & A\mathbf{1} &= 0, \\ A_{ij} &\geq 0, & A_{ij} &= 0 \iff e_{ij} \notin \mathcal{E}, \quad i \neq j. \end{aligned}$$

Define $\bar{x} = \mathbf{1}x_{\text{avg}} \in \mathbb{R}^n$ and let $M = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. A well-known result states that, under the above assumptions, the nodes will reach consensus as $t \rightarrow \infty$, i.e., $\lim_{t \rightarrow \infty} x(t) = \bar{x}$ [2]. To achieve their respective objectives, the designer and the adversary control the elements of A as we describe next. This will render the matrix A to be time-varying.

The adversary attempts to slow down convergence by breaking at most $\ell \leq m$ links at each time t . Let $u_{ij}(t) \in \{0, 1\}$ be the weight the adversary assigns to link $e_{ij} \in \mathcal{E}$ at time $t \in \mathbb{R}_{\geq 0}$. He breaks link e_{ij} when $u_{ij}(t) = 1$. Define $r := \binom{n}{2}$. The action set of the adversary is then

$$\begin{aligned} U &= \{w \in \mathbb{R}^r : w = [w_{12}, \dots, w_{1n}, w_{23}, \dots, w_{(n-1)n}]^T, w_{ij} \in \{0, 1\}, \\ &w_{ij} = 0 \text{ if } e_{ij} \notin \mathcal{E}, \|w\|_1 \leq \ell\}. \end{aligned}$$

The set of admissible controls, \mathcal{U} , consists of all functions that are piecewise continuous in time and whose range is U . Given a time interval $[0, T]$, we can formally write

$$\mathcal{U} = \{u : [0, T] \rightarrow U \mid u \text{ is a piecewise continuous function of } t\}.$$

We introduce a network designer who attempts to accelerate convergence by controlling the weights of the edges. The designer can change the weight of a given link by adding $v_{ij}(t)$ to its weight a_{ij} . We assume that $v_{ij}(t) \in \{0, b\}$ and that the number of links the designer modifies is at most $\ell \leq m$. Given the above definitions, we can write down the (i, j) -th element, $i \neq j$, of the matrix $A(u(t), v(t))$ as

$$A_{ij}(u(t), v(t)) = (a_{ij} + v_{ij}(t))(1 - u_{ij}(t)), \quad \text{for all } e_{ij} \in \mathcal{E} \quad (2)$$

We require that the resulting matrix is a negative Laplacian of the graph; hence, we must have $A_{ii}(u(t), v(t)) = -\sum_{j \neq i} A_{ij}(u(t), v(t))$, for all $i \in \mathcal{V}$.

Given a time interval $[0, T]$, define the following functional:

$$J(u, v) = \frac{1}{2} \int_0^T k(t) \|x(t) - \bar{x}\|_2^2 dt,$$

where the weighting factor $k(t)$ is positive and integrable over $[0, T]$, which can, for example, be viewed as a discounting factor, such as $k(t) = e^{-\alpha t}$ for some $\alpha > 0$. This constitutes the utility function of the adversary, and that of the designer is $-J(u, v)$. We will study two problems. In the first one, the adversary acts first by selecting the links he is interested in breaking. Then, the network designer optimizes his choices over the resulting graph, which we denote by $\mathcal{G}(u(t)) = (\mathcal{N}, \mathcal{E}(u(t)))$, where $\mathcal{E}(u(t)) = \mathcal{E} \setminus \{e_{ij} \in \mathcal{E} : u_{ij}(t) = 1\}$. In this case, the action set of the designer can be written as

$$\begin{aligned} V(u(t)) = \{ & w \in \mathbb{R}^r : w = [w_{12}, \dots, w_{1n}, w_{23}, \dots, w_{(n-1)n}]^T, w_{ij} \in \{0, b\}, \\ & w_{ij} = 0 \text{ if } e_{ij} \notin \mathcal{E}(u(t)), \|w\|_1 \leq bl \}. \end{aligned}$$

The set of admissible controls for the designer, $\mathcal{V}(u)$, consists of all piecewise continuous functions whose range is $V(u)$. Formally, we define

$$\mathcal{V}(u) = \{v : [0, T] \rightarrow V(u(t)) \mid v \text{ is a piecewise continuous function of } t\}.$$

The max–min problem can now be formally written as¹

$$\begin{aligned} & \sup_{u \in \mathcal{U}} \inf_{v \in \mathcal{V}(u)} J(u, v) \\ & \text{subject to } \dot{x}(t) = A(u(t), v(t))x(t), \quad x(0) = x_0. \end{aligned}$$

¹Even though existence of a maximum and a minimum has not yet been shown at this stage, we will still call this the “max–min” problem in anticipation of such an existence result later in the paper. The formal definition below is still in terms of sup and inf. The same argument applies to the min–max problem to be introduced shortly.

In the second problem, the order is reversed. Because the designer acts first in this problem, he can optimize over the entire graph \mathcal{G} . Thus, the action set of the designer in this problem is $V := V(0)$ and the set of its admissible controls is $\mathcal{V} := \mathcal{V}(0)$; the sets of actions and admissible controls of the adversary remain the same. We can then write

$$\begin{aligned} & \inf_{v \in \mathcal{V}} \sup_{u \in \mathcal{U}} J(u, v) \\ \text{subject to } & \dot{x}(t) = A(u(t), v(t))x(t), \quad x(0) = x_0. \end{aligned}$$

In a computer network, the max–min problem allows the network designer (who is the maximizer here) to architect networks that are robust against strategic virus diffusion. The min–max problem finds applications in army combat situations where the designer (the minimizer) attempts to counter the attacks of the enemy intending to disrupt the network communication.

Given the nature of the players’ possible modifications of the network, as described by (2), we can view the actions of the players as switches among the possible Laplacian matrices resulting from modifying the links. Moreover, the capability of the designer and the adversary to change the system matrix renders it as a “switched” one. The optimal controllers for such systems can exhibit Zeno effect, i.e., they may switch infinitely many times over a finite interval. In order to explicitly eliminate the possibility of infinite switching, we make the following assumption in the remainder of this paper.

Assumption 1. *Let $0 \leq r_1 < \dots < r_{K_u}$ be the switching times of some $u \in \mathcal{U}$ and $0 \leq s_1 < \dots < s_{K_v}$ be those of some $v \in \mathcal{V}$. We assume that $K_u, K_v \in \mathbb{Z}_{\geq 0}$ are finite, and that there exists a globally minimum dwell time $\tau > 0$ such that*

$$\tau \leq \min \{r_{i+1} - r_i, s_{i+1} - s_i, |r_i - s_j| : 1 \leq i \leq K_u, 1 \leq j < K_v\}, \quad (3)$$

over which the system matrix $A(u, v)$ is time-invariant.

Note that this assumption is well motivated for practical reasons. Consider, for example, a communication network where an adversary is a jammer injecting an interfering signal at some links. If the adversary chooses to change the set of links it is jamming, there must be some delay for the adversary to change its configuration. Now, we make the following assumption for both problems:

Assumption 2. *The initial matrix $A(0, 0)$, the time interval $[0, T]$, the values ℓ and b , and the initial state x_0 are common information to both players.*

We recall the definition of an SPE.

Definition 1 (Saddle-Point Equilibrium (SPE) [19]). *The pair (u^*, v^*) constitutes an SPE if it satisfies the following pair of inequalities*

$$J(u, v^*) \leq J(u^*, v^*) \leq J(u^*, v), \quad (4)$$

for $u \in \mathcal{U}$, $v \in \mathcal{V}$.

The following remarks are now in order.

Remark 1. *(Non-Rectangular Strategy Sets and Existence of SPE) When the strategy sets are rectangular, i.e., the strategy of one player does not restrict the strategy space of the other, the following relationship holds:*

$$\underline{V} = \sup_{u \in \mathcal{U}} \inf_{v \in \mathcal{V}} J(u, v) \leq \inf_{v \in \mathcal{V}} \sup_{u \in \mathcal{U}} J(u, v) = \overline{V}, \quad (5)$$

where $\underline{V}, \overline{V}$ are called, respectively, the lower and upper values of the game. When the strategy sets are non-rectangular, however, the order in (5) may not hold. Moreover, one should not expect the pair of inequalities (4) to hold, and hence an SPE may not exist. In the max–min problem in this paper, the strategy sets of the players are non-rectangular as the adversary’s action, removing links from \mathcal{G} , could restrict the actions available to the designer.

Remark 2. *(Problem Complexity) Let us consider the problem of the adversary for a given strategy of the designer. Assume that the adversary can act at $K_u \in \mathbb{Z}_{\geq 0}$ given time instances over the interval $[0, T]$. Then, for $\ell \leq m$, assuming that $\|u(t)\|_1 = \ell$ for all $t \in \mathbb{R}_{\geq 0}$, the total number of links that need to be tested in a brute-force approach is*

$$\binom{m}{\ell}^{K_u} \geq \left(\frac{m}{\ell}\right)^{\ell K_u}. \quad (6)$$

Clearly, the brute-force approach leads to an exponential number of computations as a function of K_u . The same argument applies to the problem faced by the network designer.

III. OPTIMAL STRATEGIES

We will now present the solutions to the two problems introduced above. In [1], we have shown that the canonical equations provided by the maximum principle (MP) are intractable due to the interdependence between the state, costate, and the optimal controls; therefore, it may not be possible to obtain the optimal strategies in closed form using the MP. Here, we take an alternative route to arrive at the optimal strategies of the players by working directly with the objective functional. In what follows, we will often drop the time index and other arguments for notational simplicity. We will be using the term ‘‘connected component’’ to refer to a set of connected nodes which have the same values. The following quantities will be central to the derivation of the optimal strategies:

$$\nu_{ij} := -(x_i - x_j)^2, \quad w_{ij} := (a_{ij} + \nu_{ij})\nu_{ij}. \quad (7)$$

A. The Min–Max Problem

The following theorem presents the optimal strategy of the adversary in the min–max problem. Define the set

$$\mathcal{L}_\ell(v) = \Phi_\ell(\{(e_{ij}, (a_{ij} + \nu_{ij})\nu_{ij}) : e_{ij} \in \mathcal{E}\}) \subseteq \mathcal{E}. \quad (8)$$

Theorem 1. *Under Assumptions 1 and 2, and for a fixed strategy v of the designer, the optimal strategy of the adversary in the min–max problem is*

$$w_{ij}^*(v) = \begin{cases} 1, & e_{ij} \in \mathcal{L}_\ell(v) \\ 0, & e_{ij} \notin \mathcal{L}_\ell(v) \end{cases}$$

If the adversary has an optimal strategy of breaking fewer than ℓ links, then either \mathcal{G} has a cut of size less than ℓ or the nodes have reached consensus by time t . In either of these cases, breaking ℓ links is also optimal.

Proof: For a fixed strategy of the designer $v \in \mathcal{V}$, we will show that it is optimal for the maximizer to rank the links based on their w_{ij} values, where w_{ij} was defined in (7). Under Assumption 1, the function x becomes piecewise continuous. Hence, the function w_{ij} , for all $e_{ij} \in \mathcal{E}$, is also piecewise continuous and its value cannot change abruptly over a finite interval. As a result, we can regard the system as a time-invariant one over a small interval $[t_0, t_0 + \delta] \subset [0, T]$, where $0 < \delta \leq \tau$, and τ was defined in (3). The proof consists of two steps.

- 1) Showing that, over a small interval $[t_0, t_0 + \delta]$, it is optimal for the adversary to switch from a strategy $u \in \mathcal{U}$ to another strategy $u^* \in \mathcal{U}$, where u^* entails breaking the ℓ links with the lowest w_{ij} values.
- 2) Showing that allowing u^* to mimic u for the remaining time of the problem preserves the gain obtained over $[t_0, t_0 + \delta]$.

Over a small interval, u and u^* induce certain system matrices. Let the system matrix corresponding to u over $[t_0, t_0 + \delta]$ be $A(u, v) = A$, and let $\|u\|_1 < \ell$ over this interval. Because the control strategies of both players are time-invariant over this interval, we have

$$x(t) = e^{A(t-t_0)}x(t_0), \quad t \in [t_0, t_0 + \delta]. \quad (9)$$

Let $P(t) := e^{At}$. Due to the structure of A , $P(t)$ is a doubly stochastic matrix for $t \geq 0$ [20, p. 63]. Note that we can write $x(t_0) = \tilde{P}x_0$, where \tilde{P} is some doubly stochastic matrix. Indeed, assume that either or both controls had switched once at some time $\tilde{t}_0 \in [0, t_0)$, and that the system matrix over $[0, \tilde{t}_0)$ was \tilde{A}_1 , and the system matrix corresponding to $[\tilde{t}_0, t_0)$ was \tilde{A}_2 . Then $x(t_0) = e^{\tilde{A}_2(t_0-\tilde{t}_0)}e^{\tilde{A}_1\tilde{t}_0}x_0$. Because both $e^{\tilde{A}_1t}$, $e^{\tilde{A}_2t}$ are doubly stochastic matrices, their product is also doubly stochastic. We can readily generalize this result to any number of switches in the interval $[0, t_0)$. With this observation, we can write

$$x(t) - \bar{x} = P(t - t_0)\tilde{P}x_0 - Mx_0 = (P(t - t_0) - M)x(t_0),$$

where the last equality follows from the fact that

$$\tilde{P}M = M\tilde{P} = M, \quad \tilde{P} \text{ is doubly stochastic.} \quad (10)$$

We want to show that switching from strategy u to strategy u^* at some time $t^* \in [t_0, t_0 + \delta]$, can improve the utility of the adversary. To this end, we assume that the matrix induced by u^* over $[t_0, t^*)$ is A , while the system matrix corresponding to u^* over $[t^*, t_0 + \delta]$ is B . Define the doubly stochastic matrix $Q(t) := e^{Bt}$, $t \geq 0$. Over $[t^*, t_0 + \delta]$, the strategies u and u^* are identical except at link $e_{ij} \in \mathcal{E}$, where $u_{ij} = 0$ and $u_{ij}^* = 1$, i.e., $\|u\|_1 < \|u^*\|_1$ over this sub-interval. It follows that:

$$A_{ij} > B_{ij} = 0, \quad A_{kl} = B_{kl} \quad \forall e_{kl} \neq e_{ij}. \quad (11)$$

Formally, we want to prove the following inequality:

$$\begin{aligned} & \int_{t_0}^{t_0+\delta} k(t) \|(P(t-t_0) - M)x(t_0)\|_2^2 dt \\ & < \int_{t_0}^{t^*} k(t) \|(P(t-t_0) - M)x(t_0)\|_2^2 dt \\ & \quad + \int_{t^*}^{t_0+\delta} k(t) \|(Q(t-t^*) - M)P(t^*-t_0)x(t_0)\|_2^2 dt, \end{aligned}$$

or equivalently

$$\begin{aligned} & \int_{t^*}^{t_0+\delta} k(t) \cdot [\|(Q(t-t^*) - M)P(t^*-t_0)x(t_0)\|_2^2 \\ & \quad - \|(P(t-t_0) - M)x(t_0)\|_2^2] dt > 0. \end{aligned} \quad (12)$$

Using (10) and the semi-group property, (12) simplifies to

$$\int_{t^*}^{t_0+\delta} k(t) \cdot x(t_0)^T \Lambda(t, t^*) x(t_0) dt > 0, \quad (13)$$

where $\Lambda(t, t^*) = P(t^*-t_0)Q(2(t-t^*))P(t^*-t_0) - P(2(t-t_0))$. A sufficient condition for (13) to hold is

$$h(t, x(t_0)) = x(t_0)^T \Lambda(t, t^*) x(t_0) > 0, \text{ for } t > t^*. \quad (14)$$

As $\delta \downarrow 0$, we can write $P(t) = I + tA + \mathcal{O}(\delta^2)$, where $\mathcal{O}(\delta^2)/\delta \leq L$ for sufficiently small δ and some finite constant L . We therefore have

$$\begin{aligned} \Lambda(t, t^*) &= (I + (t^* - t_0)A + \mathcal{O}(\delta^2)) (I + 2(t - t^*)B \\ & \quad + \mathcal{O}(\delta^2)) (I + (t^* - t_0)A + \mathcal{O}(\delta^2)) - (I + 2(t - t_0)A \\ & \quad + \mathcal{O}(\delta^2)) = 2(t - t^*)B + 2(t^* - t_0)A - 2(t - t_0)A \\ & \quad + \mathcal{O}(\delta^2) = 2(t - t^*)(B - A) + \mathcal{O}(\delta^2). \end{aligned} \quad (15)$$

For sufficiently small δ , the first term dominates the second term. Recall that the quadratic form of a Laplacian matrix L exhibits the following form: $x^T L x = \sum_{l=1}^n \sum_{k=1}^{l-1} L_{kl} (x_l - x_k)^2$, for any $x \in \mathbb{R}^n$. Note that $B - A$ is in fact a negative Laplacian. Using (11), we can then write

$$\begin{aligned} h(t, x(t_0)) &= 2(t - t^*) \sum_{r>s} (A_{sr} - B_{sr}) (x_r(t_0) - x_s(t_0))^2 + \mathcal{O}(\delta^2) \\ &= 2(t - t^*) A_{ij} (x_j(t_0) - x_i(t_0))^2 + \mathcal{O}(\delta^2). \end{aligned} \quad (16)$$

For small enough δ , the higher order terms are dominated by the first term. Hence, if there is a link e_{ij} such that $x_i(t_0) \neq x_j(t_0)$, there exists t^* such that $h(t, x(t_0)) > 0$ for $t \in (t^*, t_0 + \delta]$. Since t_0 was arbitrary, we conclude that the optimal strategy must satisfy $\|u^*(t)\|_1 = \ell$ for all t , given that each of the ℓ links connects two nodes having different values.

If no link such that $x_i(t_0) \neq x_j(t_0)$ exists at a given time t_0 , the adversary does not need to break additional links, although breaking more links does not affect optimality because $h(t, x(t_0)) = 0$ in such a case. There are two cases under which the adversary cannot find a link to make $h(t, x(t_0)) > 0$: (i) The graph at time t_0 is one connected component. In this case, the nodes have already reached consensus and $\|u^*(t)\|_1 < \ell$. This is a *losing strategy* for the adversary as he has failed in preventing nodes from reaching agreement; (ii) The graph at time t_0 has multiple connected components, and the number of links connecting the components is less than ℓ . The adversary here possesses a *winning strategy* with $\|u^*(t)\|_1 < \ell$, as he can disconnect \mathcal{G} into multiple components and prevent consensus.

Next, we need to show that the adversary will modify the ℓ links with the lowest w_{ij} values. Let us again restrict our attention to the interval $[t_0, t_0 + \delta]$ where the adversary applies strategy u . Assume (to the contrary) that the links the adversary breaks over this interval are not the ones with the lowest w_{ij} values. In particular, assume that the adversary chooses to break link e_{kl} , while there is a link e_{ij} such that $w_{ij} < w_{kl}$. Assume that the adversary switches at time $t^* \in [t_0, t_0 + \delta]$ to strategy u^* by *breaking* link e_{ij} and *unbreaking* link e_{kl} . Then, (16) becomes

$$h(t, x(t_0)) = 2(t - t^*) (w_{kl}(t_0) - w_{ij}(t_0)) + \mathcal{O}(\delta^2).$$

Hence, by following the same arguments as above, we can conclude that breaking e_{kl} is not optimal.

The second step of the proof is to show that switching to strategy u^* guarantees an improved utility for the adversary *regardless of how the original trajectory corresponding to u changes beyond time $t_0 + \delta$* . To this end, we will assume that from time $t_0 + \delta$ onward, strategy u^* will *mimic* strategy u . Assume that strategy u switches from matrix A to matrix C over the interval $[t_0 + \delta, t_0 + 2\delta]$, and define $R(t) := e^{Ct}$. Hence, strategy u^* will also switch from the system matrix B to matrix C . However, the trajectories corresponding to u and u^* will have different initial conditions at time $t_0 + \delta$, due to the switch that strategy u^* made at time t^* . Fig. 1 illustrates this idea. Recall that according to A , we have $\|u\|_1 < \ell$ and $u_{ij} = 0$. Here, the

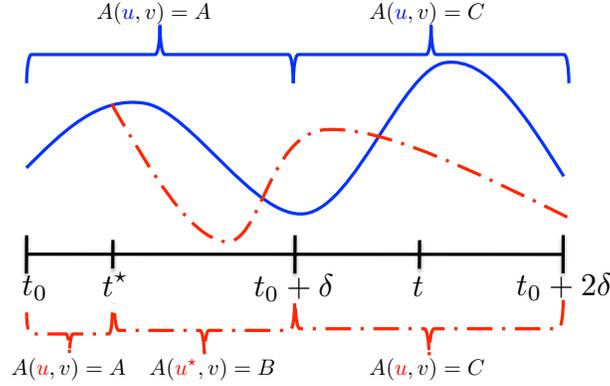


Fig. 1. A demonstration of the technique used in the proof. The blue solid trajectory corresponds to u while the red dashed trajectory corresponds to u^* .

system matrix B can differ from the matrix A in two ways: either (i) B dictates breaking one additional link compared to A , or (ii) B dictates breaking link e_{ij} and unbreaking link e_{kl} where $w_{ij} < w_{kl}$. Consider Case (i) first and let us study the behavior of the system over the interval $[t_0 + \delta, t_0 + 2\delta]$ where we can assume that the system is time-invariant. To show that the gain obtained over $[t_0, t_0 + \delta]$ by the switch made by u^* is maintained over $[t_0 + \delta, t_0 + 2\delta]$, we must prove the following inequality:

$$\int_{t_0 + \delta}^{t_0 + 2\delta} k(t) \cdot [L_1 - L_2] dt > 0, \quad (17)$$

where

$$L_1 := \|(R(t - (t_0 + \delta)) - M)Q(t_0 + \delta - t^*)P(t^* - t_0)x(t_0)\|_2^2,$$

$$L_2 := \|(R(t - (t_0 + \delta)) - M)P(t_0 + \delta - t_0)x(t_0)\|_2^2.$$

As before, it suffices to prove that the integrand $L_1 - L_2$ is positive. Let us now expand both L_1 and L_2 .

$$\begin{aligned} L_1 &= x(t_0)^T P(t^* - t_0)Q(t_0 + \delta - t^*)(R(t - (t_0 + \delta)) - M)(R(t - (t_0 + \delta)) - M) \\ &\quad Q(t_0 + \delta - t^*)P(t^* - t_0)x(t_0) \\ &= x(t_0)^T (P(t^* - t_0)Q(t_0 + \delta - t^*)R(2(t - (t_0 + \delta))))Q(t_0 + \delta - t^*)P(t^* - t_0) - M)x(t_0). \end{aligned}$$

Similarly, $L_2 = x(t_0)^T(P(\delta)R(2(t - (t_0 + \delta))))P(\delta) - M)x(t_0)$. We can then write

$$\begin{aligned} L_1 - L_2 &= x(t_0)^T(P(t^* - t_0)Q(t_0 + \delta - t^*)R(2(t - (t_0 + \delta))))Q(t_0 + \delta - t^*)P(t^* - t_0) \\ &\quad - P(\delta)R(2(t - (t_0 + \delta)))P(\delta)x(t_0) \\ &:= x(t_0)^T(F_1 - F_2)x(t_0). \end{aligned}$$

Before we perform a first-order Taylor expansion to the above terms, let us define the following quantities: $\tau_1 = t^* - t_0$, $\tau_2 = (t_0 + \delta) - t^*$, and $\tau_3 = t - (t_0 + \delta)$, where $t^* \in [t_0, t_0 + \delta]$ and $t \in [t_0 + \delta, t_0 + 2\delta]$. Using Proposition 1 in the Appendix, we can now expand F_1 and F_2 as follows:

$$\begin{aligned} F_1 &= (I + \tau_1 A + \mathcal{O}(\tau_1^2)) (I + \tau_2 B + \mathcal{O}(\tau_2^2)) (I + 2\tau_3 C + \mathcal{O}(\tau_3^2)) (I + \tau_2 B + \mathcal{O}(\tau_2^2)) \\ &\quad (I + \tau_1 A + \mathcal{O}(\tau_1^2)) \\ &= I + 2\tau_1 A + 2\tau_2 B + 2\tau_3 C + \mathcal{O}(\delta^2) \\ F_2 &= (I + \delta A + \mathcal{O}(\delta^2)) (I + 2\tau_3 C + \mathcal{O}(\tau_3^2)) (I + \delta A + \mathcal{O}(\delta^2)) \\ &= I + 2\delta A + 2\tau_3 C + \mathcal{O}(\delta^2). \end{aligned}$$

Hence, we have $F_1 - F_2 = 2((t_0 + \delta) - t^*)(B - A) + \mathcal{O}(\delta^2)$, and thereby we obtain

$$\begin{aligned} L_1 - L_2 &= 2((t_0 + \delta) - t^*) \sum_{r>s} (A_{sr} - B_{sr}) (x_r(t_0) - x_s(t_0))^2 + \mathcal{O}(\delta^2) \\ &= 2(t_0 + \delta - t^*) A_{ij} (x_j(t_0) - x_i(t_0))^2 + \mathcal{O}(\delta^2). \end{aligned}$$

If instead the matrix B dictates breaking link e_{ij} and unbreaking link e_{kl} , where $w_{ij} < w_{kl}$, the difference in the utilities would be $L_1 - L_2 = 2(t_0 + \delta - t^*)(w_{kl}(t_0) - w_{ij}(t_0)) + \mathcal{O}(\delta^2)$. Hence, in both cases, for small enough δ , we conclude that $L_1 - L_2 > 0$, which implies that (17) is satisfied, and the gain obtained by switching to system matrix B at $t^* \in [t_0, t_0 + \delta]$ is maintained over $[t_0 + \delta, t_0 + 2\delta]$. Note that the effect of switching to matrix C is cancelled out in $F_1 - F_2$, and hence $L_1 - L_2$, since the strategy u^* is mimicking strategy u . Hence, by partitioning the interval $(t_0 + 2\delta, T]$ into small sub-intervals of length δ and repeating the above analysis, we conclude that the gain due to the switch at time t^* is preserved over the remaining time of the problem. \blacksquare

We can now derive the optimal strategy of the designer in the min-max problem. Recall the set $\mathcal{L}_\ell(v) \subseteq \mathcal{E}$ defined in (8). Let $\mathcal{L}_{\ell,k}(v) \in \mathcal{E}$ denote the k -th link of $\mathcal{L}_\ell(v)$, $k \in \{1, \dots, \ell\}$.

TABLE I

ALGORITHM I: COMPUTING THE OPTIMAL STRATEGY FOR THE MINIMIZER IN THE MIN-MAX PROBLEM.

```

0: input: a strategy  $v$  with  $\|v\|_1 = 0$ 
1: for  $i = \ell \downarrow 1$ 
2:   if  $\exists S \subseteq \Phi(\mathcal{P}(0)), |S| = i, \mathcal{L}_{\ell,i}(0) \notin \mathcal{L}_\ell(v_S(b))$ 
3:     Set  $v_{ij}^* = b, \forall e_{ij} \in S \cup \Phi_{\ell-i}(\overline{\mathcal{P}}(v_S(b)))$ .
4:     Exit for loop.
5:   end
6: end
7: if  $\|v\|_1 = 0$ 
8:   Set  $v_{ij}^* = b$  for all  $e_{ij} \in \Phi_\ell(\overline{\mathcal{P}}(0))$ .
9: end

```

Also, define $\mathcal{L}_{\ell,k}^{-1}(v) \in \mathbb{R}$ as the value such that $\Phi(\mathcal{L}_{\ell,k}(v), \mathcal{L}_{\ell,k}^{-1}(v)) = \mathcal{L}_{\ell,k}(v)$. We assume that $\mathcal{L}_{\ell,1}^{-1}(v) \geq \dots \geq \mathcal{L}_{\ell,\ell}^{-1}(v)$. Further, define the sets $\mathcal{P}(v) = \{(e_{ij}, a_{ij}\nu_{ij}) : e_{ij} \notin \mathcal{L}_\ell(v)\} \subset \mathcal{E} \times \mathbb{R}$ and $\overline{\mathcal{P}}(v) = \{(e_{ij}, \nu_{ij}) : e_{ij} \notin \mathcal{L}_\ell(v)\} \subset \mathcal{E} \times \mathbb{R}$. We also define

$$[v_S(b)]_{ij} = \begin{cases} b, & e_{ij} \in S \\ 0, & e_{ij} \notin S \end{cases}$$

Theorem 2. *In the min-max problem, and under Assumptions 1 and 2, the optimal strategy of the designer is to run Algorithm I and set $v_{ij}^* \in \{0, b\}$ if $\nu_{ij} = 0$. Further, it is optimal for the designer to modify ℓ links.*

Proof: By Theorem 1, we deduce that $\|v^*(t)\|_1 = bl$, because the designer would be at a disadvantage if he modifies fewer links than the adversary.

We first consider the designer's strategy over a fixed small interval $[t_0, t_0 + \delta]$ over which both u and v are fixed. Using similar steps as those leading to (13), and after applying a first-order Taylor expansion, we can write the designer's utility over $[t_0, t_0 + \delta]$ as

$$\int_{t_0}^{t_0+\delta} k(t) \cdot 2(t - t_0) \sum_{j>i} (a_{ij} + v_{ij})(1 - u_{ij})(x_i(t_0) - x_j(t_0))^2 dt + \mathcal{O}(\delta^2). \quad (18)$$

According to Theorem 1, and in the absence of the designer, it is optimal for the adversary to break the links in $\mathcal{L}_\ell(0)$. Therefore, the designer must attempt to modify the ranking of the links such that the links (or a subset of them) in $\mathcal{L}_\ell(0)$ are not in $\mathcal{L}_\ell(v^*)$. In essence, this is what

Algorithm I attempts to achieve. Being of the lowest negative value, and hence the link both the adversary and the designer are interested in, let us explore how the designer can push $\mathcal{L}_{\ell,\ell}(0)$ higher in the ranking of the link values. The designer can achieve this if under some strategy $v \in \mathcal{V}$, the value $\mathcal{L}_{\ell,\ell}^{-1}(0)$ is no longer among the lowest ℓ negative values; in other words, the designer can alter the ranking if there is a set $\mathcal{S} \subset \mathcal{P}(0)$, $|\mathcal{S}| = \ell$, such that when he sets $v_{ij} = b$ for all links in \mathcal{S} , there will be ℓ values that are smaller than $\mathcal{L}_{\ell,\ell}^{-1}(0)$ (steps 2 and 3 in Algorithm I). The adversary will then break the links in \mathcal{S} and will spare the link corresponding to $\mathcal{L}_{\ell,\ell}^{-1}(0)$ as required. To see why this is optimal, consider the following two cases, covering the types of links that can be in \mathcal{S} .

Case 1: If a link in \mathcal{S} is also in $\mathcal{L}_\ell(0)$, then this is optimal due to the fact that the adversary will disconnect that link since it is in $\mathcal{L}_\ell(0)$. Hence, if the designer can utilize this link to modify the ranking and protect a link whose associated value is more negative ($\mathcal{L}_{\ell,\ell}(0)$ in this case), then this can only improve his utility. The same reasoning applies if more than one of the links in \mathcal{S} are also in $\mathcal{L}_\ell(0)$.

Case 2: If none of the links in \mathcal{S} is in $\mathcal{L}_\ell(0)$, then necessarily some of the links in $\mathcal{L}_\ell(0)$ will also be protected along with the link corresponding to $\mathcal{L}_{\ell,\ell}^{-1}(0)$. This is because $|\mathcal{S}| = \ell$, and the adversary can break at most ℓ links. Hence, this scenario is more favorable to the designer than the previous one and can therefore only improve his utility.

If such an \mathcal{S} exists, then the designer would have exhausted all possible moves, since $|\mathcal{S}| = \ell$, and the algorithm terminates (step 4 of the algorithm). Otherwise, if no such set exists in $\mathcal{P}(0)$, then the designer should try to protect the next most negative link whose value is precisely $\mathcal{L}_{\ell,\ell-1}^{-1}(0)$ by finding a set \mathcal{S} of size $\ell - 1$. Since $\mathcal{L}_{\ell,\ell-1}^{-1}(0) \geq \mathcal{L}_{\ell,\ell}^{-1}(0)$, the link corresponding to $\mathcal{L}_{\ell,\ell}^{-1}(0)$ along with \mathcal{S} will constitute the set of ℓ links that the adversary will break. Then, the designer should set $v_{ij} = b$ for all the links in \mathcal{S} , and for the remaining action the designer should select the link with the most negative v_{ij} that is *not* in $\mathcal{L}_\ell(v_{\mathcal{S}}(b))$; this is precisely the set $\Phi_1(\overline{\mathcal{P}}(v_{\mathcal{S}}(b)))$ (step 3 of the algorithm). The reason behind searching in $\overline{\mathcal{P}}(v_{\mathcal{S}}(b))$ and not in $\mathcal{P}(v_{\mathcal{S}}(b))$ after finding \mathcal{S} is that the a_{ij} 's only affect the utility of the designer when he attempts to alter the ranking.

This procedure then repeats until the designer has tried to protect all the links in $\mathcal{L}_\ell(0)$. If the designer fails in protecting *all* the links in $\mathcal{L}_\ell(0)$, then we must have $\|v\|_1 = 0$, i.e., the input

strategy was not altered. Then, the optimal strategy is to set $v_{ij} = b$ for the links with most negative ν_{ij} 's in $\overline{\mathcal{P}}(0)$ (steps 7 and 8 in Algorithm I).

The final step of the proof is to show that applying Algorithm I over $[0, T]$ is optimal for the designer. To this end, it suffices to show that modifying links with lower ν_{ij} values is more beneficial to the designer, as Algorithm I attempts to protect these links. Given the links $e_{ij}, e_{kl} \in \mathcal{E}$, assume that $\nu_{ij} < \nu_{kl}$. Consider the two system matrices A and B , and let $v_{ij} = 0$, $v_{kl} = b$ and $v_{ij}^* = b$, $v_{kl}^* = 0$. Assume that a strategy v dictates applying matrix A over $[t_0, t_0 + \delta]$ and applying the matrix C over $[t_0 + \delta, t_0 + 2\delta]$. Also, assume that according to v^* , the designer applies A over $[t_0 + \delta, t^*]$, B over $(t^*, t_0 + \delta]$, and C over $[t_0 + \delta, t_0 + 2\delta]$. Following the steps presented in step 2 of the proof of Theorem 1, we conclude that, for δ small enough, the quantity $-2(t_0 + \delta - t^*)b(\nu_{kl} - \nu_{ij}) + \mathcal{O}(\delta^2)$ is negative. It then follows that the gain obtained by switching to system matrix B at $t^* \in [t_0, t_0 + \delta]$ is maintained over $[t_0 + \delta, t_0 + 2\delta]$. Hence, by partitioning the interval $(t_0 + 2\delta, T]$ into small sub-intervals of length δ and repeating the above analysis, we conclude that Algorithm I is optimal over $[0, T]$. ■

B. The Max–Min Problem

The following theorem specifies the optimal strategies of the adversary and the designer in the max–min problem. Let $\mathcal{F}_\ell(u) = \Phi_\ell(\{(e_{ij}, \nu_{ij}) : e_{ij} \in \mathcal{E}(u)\}) \subset \mathcal{E}(u)$, where we recall that $\mathcal{E}(u) = \mathcal{E} \setminus \{e_{ij} \in \mathcal{E} : u_{ij}(t) = 1\}$, for some $u \in \mathcal{U}$. If $m < 2\ell$, the sets $\mathcal{E}(u), \mathcal{F}_\ell(u)$ could contain fewer than ℓ links. For simplicity, we assume that $m \geq \ell$ in the following proof, which guarantees that $|\mathcal{F}_\ell(u)| = \ell$. However, the result of the theorem applies regardless of this assumption, and the modification of the proof is straightforward.

Theorem 3. *Under Assumptions 1 and 2, and for a fixed strategy u of the adversary, the optimal strategy of the network designer in the max–min problem is given by*

$$v_{ij}^*(u) = \begin{cases} b, & e_{ij} \in \mathcal{F}_\ell(u) \\ 0, & e_{ij} \notin \mathcal{F}_\ell(u) \end{cases}$$

If the designer has an optimal strategy of modifying fewer than ℓ links, then either \mathcal{G} has a cut of size less than ℓ or the nodes have reached consensus by time t . In either of these cases, breaking ℓ links is also optimal.

Proof: The proof follows the same two steps used to prove Theorem 1. For a fixed strategy of the adversary u , we will show that it is optimal for the minimizer to rank the links based on their ν_{ij} values. Under Assumption 1, the function x becomes piecewise continuous. Hence, the function ν_{ij} , for all $e_{ij} \in \mathcal{E}(u)$, is also piecewise continuous and its value cannot change abruptly over a finite interval. As a result, we can regard the system as a time-invariant one over a small interval $[t_0, t_0 + \delta] \subset [0, T]$, where $0 < \delta \leq \tau$, and τ was defined in (3).

Let v be an arbitrary strategy of the designer with $\|v\|_1 < b\ell$. Over a small interval, v and v^* induce certain system matrices. Let the system matrix corresponding to v over $[t_0, t_0 + \delta]$ be $A(u, v) = A$. Because the control strategies of both players are time-invariant over this interval, the state trajectory is given by (9). We want to show that switching from strategy v to strategy v^* at some time $t^* \in [t_0, t_0 + \delta]$ can improve the utility of the designer. To this end, we assume that the matrix induced by v^* over $[t_0, t^*]$ is A , while the system matrix corresponding to v^* over $[t^*, t_0 + \delta]$ is B . Assume that $e_{ij} \in \mathcal{E}(u)$, i.e., $u_{ij} = 0$. Over $[t^*, t_0 + \delta]$, the strategies v and v^* are identical except at link e_{ij} , where $v_{ij} = 0$ and $v_{ij}^* = b$, i.e., $\|v\|_1 < \|v^*\|_1$ over this sub-interval. It follows that:

$$B_{ij} = a_{ij} + b > A_{ij} = a_{ij}, \quad A_{kl} = B_{kl}, \quad \forall e_{kl} \neq e_{ij}. \quad (19)$$

Following similar steps to those in the proof of Theorem 1, we conclude that it suffices to prove

$$h(t, x(t_0)) = x(t_0)^T \Lambda(t, t^*) x(t_0) < 0, \quad \text{for } t > t^*,$$

where $\Lambda(t, t^*)$ was defined in the proof of Theorem 1. For sufficiently small δ , we can arrive at the expansion in (15). Using (19) and properties of Laplacian matrices, we can then write

$$\begin{aligned} h(t, x(t_0)) &= 2(t - t^*) \sum_{r>s} (A_{sr} - B_{sr}) (x_r(t_0) - x_s(t_0))^2 + \mathcal{O}(\delta^2) \\ &= -2(t - t^*)b (x_j(t_0) - x_i(t_0))^2 + \mathcal{O}(\delta^2). \end{aligned} \quad (20)$$

For small enough δ , the higher order terms are dominated by the first term. Hence, if there is a link e_{ij} such that $x_i(t_0) \neq x_j(t_0)$, there exists t^* such that $h(t, x(t_0)) < 0$ for $t \in (t^*, t_0 + \delta]$. Since t_0 was arbitrary, we conclude that the optimal strategy must satisfy $\|v^*(t)\|_1 = b\ell$ for all t , given that each of the ℓ links connects two nodes having different values.

If no link such that $x_i(t_0) \neq x_j(t_0)$ exists at a given time t_0 , the designer does not need to break additional links, although breaking more links does not affect optimality because $h(t, x(t_0)) = 0$

in such a case. There are two cases where the designer cannot find a link to make $h(t, x(t_0)) < 0$, and they were presented in the proof of Theorem 1 in the case of the adversary. However, unlike the case of the adversary, Case (i) presents a winning strategy for the designer as the nodes are in agreement. Case (ii) is not necessarily a winning or a losing strategy for the designer.

Next, we need to show that the designer will modify the ℓ links in $\mathcal{E}(u)$ with the lowest ν_{ij} values. Let us again restrict our attention to the interval $[t_0, t_0 + \delta]$ where the designer applies strategy v . Assume (to the contrary) that the links the designer modifies over this interval are not the ones with the lowest ν_{ij} values. In particular, assume that the designer chooses to modify link $e_{kl} \in \mathcal{E}(u)$, while there is a link $e_{ij} \in \mathcal{E}(u)$ such that $\nu_{ij} < \nu_{kl}$. Assume that the designer switches at time $t^* \in [t_0, t_0 + \delta]$ to strategy v^* by modifying link e_{ij} instead of link e_{kl} . Then, (20) becomes

$$h(t, x(t_0)) = -2(t - t^*)b(\nu_{kl}(t_0) - \nu_{ij}(t_0)) + \mathcal{O}(\delta^2).$$

Hence, by following the same arguments as above, we can conclude that modifying e_{kl} is not optimal.

The second step of the proof is to show that switching to strategy v^* guarantees an improved utility for the designer regardless of how the original trajectory corresponding to v changes beyond time $t_0 + \delta$. To this end, we will assume that from time $t_0 + \delta$ onward, strategy v^* will mimic strategy v . Assume that strategy v switches from matrix A to matrix C over the interval $[t_0 + \delta, t_0 + 2\delta]$. Hence, strategy v^* will also switch from the system matrix B to matrix C . However, the trajectories corresponding to v and v^* will have different initial conditions at time $t_0 + \delta$, due to the switch that strategy v^* made at time t^* . Recall that according to A , we have $\|v\|_1 < b\ell$ and $v_{ij} = 0$. Here, the system matrix B can differ from the matrix A in two ways: either (i) B dictates modifying one additional link compared to A , or (ii) B dictates modifying link e_{ij} instead of link e_{kl} where $\nu_{ij} < \nu_{kl}$. Consider the behavior of the system over the interval $[t_0 + \delta, t_0 + 2\delta]$ where we can assume that the system is time-invariant. To show that the gain obtained over $[t_0, t_0 + \delta]$ by the switch made by v^* is maintained over $[t_0 + \delta, t_0 + 2\delta]$, it suffices to prove that the integrand $L_1 - L_2$ is negative, where L_1 and L_2 were defined in the proof of Theorem 1. For Case (i), by following the steps presented in the proof of Theorem 1, we can write

$$L_1 - L_2 = -2(t_0 + \delta - t^*)b(x_j(t_0) - x_i(t_0))^2 + \mathcal{O}(\delta^2).$$

For Case (ii), the difference in utilities would be

$$L_1 - L_2 = 2(t_0 + \delta - t^*)(w_{kl}(t_0) - w_{ij}(t_0)) + \mathcal{O}(\delta^2).$$

Hence, for small enough δ , we conclude that $L_1 - L_2 < 0$. By partitioning the interval $(t_0 + 2\delta, T]$ into small sub-intervals of length δ and repeating the above analysis, we conclude that the gain due to the switch at time t^* is preserved over the remaining time of the problem. This concludes the proof. \blacksquare

Next, we present the optimal strategy of the adversary. To this end, define the set

$$\mathcal{D}_\ell = \Phi_\ell(\{(e_{ij}, a_{ij}\nu_{ij}) : e_{ij} \in \mathcal{E}\} \cup \{(e_{ij}, (a_{ij} + b)\nu_{ij}) : e_{ij} \in \mathcal{E}\}).$$

Theorem 4. *In the max–min problem, and under Assumptions 1 and 2, the optimal strategy of the adversary is given by*

$$u_{ij}^*(t) = \begin{cases} 1, & e_{ij} \in \mathcal{D}_\ell \\ 0, & e_{ij} \notin \mathcal{D}_\ell \end{cases}$$

Further, it is optimal for the adversary to break ℓ links.

Proof: By Theorem 3, we deduce that $\|u^*(t)\|_1 = \ell$, because the adversary would be at a disadvantage if he breaks fewer links than the designer. We first consider the adversary's strategy over a fixed small interval $[t_0, t_0 + \delta]$ over which both u and v are fixed. Using a first-order Taylor expansion, the adversary's utility over $[t_0, t_0 + \delta]$ is given by (18).

In this problem, the adversary has the first-mover-advantage and needs to dispose of the links that can reduce his utility. The adversary knows that, according to $v^*(u)$, the designer attempts to make the ν_{ij} 's smaller by adding b to the corresponding edge weights. However, we cannot rule out the possibility that $(a_{lk} + b)\nu_{lk} > a_{ij}\nu_{ij}$, for some links e_{kl} and e_{ij} . Hence, the adversary is not only interested in finding the smallest negative $(a_{ij} + b)\nu_{ij}$'s, but also needs to consider the $a_{ij}\nu_{ij}$'s themselves. It follows that the adversary needs to find the terms that can become very small (negative) and set $u_{ij} = 1$ to the corresponding links. But those links are exactly the ones included in \mathcal{D}_ℓ . Formally, we can write

$$-\sum_{\substack{j>i \\ e_{ij} \in \mathcal{D}_\ell}} (a_{ij} + v_{ij})\nu_{ij} \leq -\sum_{\substack{j>i \\ e_{ij} \notin \mathcal{D}_\ell}} (a_{ij} + v_{ij})\nu_{ij},$$

This confirms that, over the interval $[t_0, t_0 + \delta]$, u^* is as claimed.

The final step of the proof is to show that switching from a strategy u to strategy u^* guarantees an improved utility for the designer over $[0, T]$. To this end, it suffices to show that modifying links with lower w_{ij} values is more beneficial to the adversary. For the links $e_{ij}, e_{kl} \in \mathcal{E}$, assume that $w_{ij} < w_{kl}$. Consider the two system matrices A and B , and let $u_{ij} = 0$, $u_{kl} = 1$ and $u_{ij}^* = 1$, $u_{kl}^* = 0$. Assume that the strategy u dictates applying matrix A over $[t_0, t_0 + \delta]$ and applying the matrix C over $[t_0 + \delta, t_0 + 2\delta]$. On the other hand, we assume that according to u^* , the adversary applies A over $[t_0 + \delta, t^*]$, B over $(t^*, t_0 + \delta]$, and C over $[t_0 + \delta, t_0 + 2\delta]$. Following the steps presented in step 2 of the proof of Theorem 1, we conclude that, for δ small enough, the quantity $2(t_0 + \delta - t^*)(w_{kl} - w_{ij}) + \mathcal{O}(\delta^2)$ is positive, which implies that the gain obtained by switching to system matrix B at $t^* \in [t_0, t_0 + \delta]$ is maintained over $[t_0 + \delta, t_0 + 2\delta]$. Hence, by partitioning the interval $(t_0 + 2\delta, T]$ into small sub-intervals of length δ and repeating the above analysis, we conclude that u^* is optimal over $[0, T]$. ■

Remark 3. (*Potential-Theoretic Analogy*) When the graph is viewed as an electrical network, $a_{ij} + v_{ij}$ can be viewed as the conductance of link $e_{ij} \in \mathcal{E}$, and $x_i - x_j$ as the potential difference across the link. Therefore, according to Theorems 2 and 3, the optimal strategy of the designer in both problems involves finding the links with the highest potential difference (or the lowest v_{ij} 's) and increasing the conductance of those links by setting $v_{ij} = b$. This leads to increasing the power dissipation across those links, which translates to increasing the information flow across the network and results in faster convergence. The optimal strategy of the adversary should therefore involve breaking the links with the highest power dissipation. But power dissipation is given by $(a_{ij} + v_{ij})(x_i - x_j)^2$, and this is exactly what the adversary targets according to Theorems 1 and 4.

C. From Potential Theory to the Maximum Principle

In this section, we show that the strategies derived in the above theorems satisfy the first-order necessary conditions for optimality given by the maximum principle (MP). We will address here the min–max problem; a theorem similar to the one presented below can be obtained also for the max–min problem. In [1], we showed that the optimal strategies provided by the MP for the min–max problem are the same as those derived in Theorems 1 and 2, with the ranking of the links performed after replacing the quantity v_{ij} with the quantity $(p_j - p_i)(x_i - x_j)$, where p is

the costate vector. The next theorem states that the potential-theoretic strategies satisfy the MP if the controllers do not switch infinitely many times over $[0, T]$.

Theorem 5. *Under Assumptions 1 and 2, the optimal strategies in Theorems 1 and 2 satisfy the canonical equations of the MP.*

Proof: See the Appendix. ■

D. Complexity of the Optimal Strategies

We next study the complexity of the optimal strategies. We first start with the max–min problem. Assuming, as in Remark 2, that the players switch their strategies a total of K times over $[0, T]$, we conclude that the worst-case complexity of the strategy of either player is $\mathcal{O}(K \cdot m \log m)$ as their strategies involve merely the ranking of sets of size at most $2m$. As for the min–max problem, the complexity of the adversary’s strategy is $\mathcal{O}(K \cdot m \log m)$. The main bottleneck in the strategy of the designer is step 2 in Algorithm I. The size of the set $\mathcal{P}(0)$ is at most $m - \ell$; thus, the worst-case complexity for the designer is $K \cdot \sum_{i=1}^{m-\ell} \binom{m-\ell}{i} \approx K \cdot \sum_{i=1}^{\ell} (m - \ell)^i$. By comparison with (6), we conclude that the optimal strategies achieve vast complexity reductions.

E. An Illustrative Example

The goal of this example is twofold: (i) to show how the players execute their strategies; and (ii) to serve as a counter example showing that an SPE may not exist and to provide some guidelines as to when one would exist. We will study the interaction between the designer and the adversary for the case when $T = \tau$, and τ is very small. By Assumption 1, we conclude that the players cannot change the actions they choose at time $t = 0$. Assume that \mathcal{G} is a complete graph with three nodes with the following weights:

$$A(0, 0) = \begin{bmatrix} -4 & 3 & 1 \\ 3 & -5 & 2 \\ 1 & 2 & -3 \end{bmatrix}.$$

Define $e_1 = (1, 2)$, $e_2 = (2, 3)$, $e_3 = (1, 3)$. Let $\nu_{12} = -1$, $\nu_{23} = -2$, and $\nu_{13} = -5$. Let $x(0) = [1, 2, 3]^T$ and $\ell = 1$. Consider the following two cases:

Case 1: ($b = 1$) Let us first consider the max–min problem. We have $\mathcal{D}_1 = \Phi_1(\{(e_1, -3), (e_1, -4), (e_2, -4), (e_3, -5), (e_2, -6), (e_3, -10)\}) = \{e_3\}$. Hence, according to Theorem 4, the adversary

breaks e_3 , and we have that $\mathcal{E}(u^*) = \mathcal{E} \setminus e_3$. We also have $\mathcal{F}_1(u^*) = \{e_2\}$, which means that $v^* = [0, 1, 0]^T$ and $u^* = [0, 0, 1]^T$. Hence, using (18), we can write

$$\begin{aligned} \underline{V} &= \int_0^T k(t) \cdot 2t[3(x_1(0) - x_2(0))^2 + 3(x_2(0) - x_3(0))^2]dt + \mathcal{O}(\delta^2) \\ &= \int_0^T k(t) \cdot 12tdt + \mathcal{O}(\delta^2). \end{aligned}$$

For the min–max problem, Algorithm I uses the following sets $\mathcal{L}_1(0) = \{e_3\}$ and $\mathcal{P}(0) = \{(e_1, -3), (e_2, -4)\}$. Let $\mathcal{S} = \{e_2\}$, and note that $\mathcal{S} \in \Phi(\mathcal{P}(0)) = \{e_1, e_2\}$. We then have $v_{\mathcal{S}}(1) = [0, 1, 0]^T$ and $\mathcal{L}_1(v_{\mathcal{S}}(1)) = \{e_2\}$. Note that $\mathcal{L}_1(0) \notin \mathcal{L}_1(v_{\mathcal{S}}(1))$. Hence, the condition in step 2 of the algorithm is satisfied with this choice of \mathcal{S} , and we have $v^* = v_{\mathcal{S}}(1)$. Then, Theorem 2 says that the designer will increase the weight of e_2 , and Theorem 1 says that the adversary will break the same link, i.e., $v^* = [0, 1, 0]^T$ and $u^* = [0, 1, 0]^T$. We thus have

$$\bar{V} = \int_0^T k(t) \cdot 14tdt + \mathcal{O}(\delta^2).$$

We conclude that in this case $\bar{V} > \underline{V}$, and an SPE does not exist.

Case 2: ($b = 0.4$) By repeating the above steps, we conclude that in the max–min problem we have $v^* = [0, 0.4, 0]^T$ and $u^* = [0, 0, 1]^T$, and we can write

$$\underline{V} = \int_0^T k(t) \cdot 10.8tdt + \mathcal{O}(\delta^2).$$

For the min–max problem, one cannot find a set \mathcal{S} satisfying the conditions of step 2 in Algorithm I. To execute step 8 of the algorithm, note that $\mathcal{L}_1(0) = \{e_3\}$, and hence $\Phi_1(\bar{\mathcal{P}}(0)) = \{e_2\}$. We therefore have $v^* = [0, 0.4, 0]^T$ and $u^* = [0, 0, 1]^T$, and hence

$$\bar{V} = \int_0^T k(t) \cdot 10.8tdt + \mathcal{O}(\delta^2).$$

In this case, the pair of inequalities (4) are satisfied and an SPE exists. The main difference between the two cases was that the designer was able to find a set \mathcal{S} that allows him to alter the ranking and deceive the adversary when $b = 1$. This made the adversary break e_3 in the max–min problem and break e_2 in the min–max problem which led to having $\bar{V} \neq \underline{V}$. When such a set does not exist, the strategy of the adversary is unchanged in both problems, and hence the upper and lower values would agree. Hence, for an SPE to exist, one needs a behavior similar to Case 2 to occur throughout the problem horizon $[0, T]$. This of course depends on the value of b and the weights a_{ij} . Section IV explores the question of existence of an SPE further.

IV. A SUFFICIENT CONDITION FOR THE EXISTENCE OF AN SPE

Thus far, we have solved the min–max and max–min problems separately and showed that the derived optimal strategies achieve the upper and lower values. Hence, to prove the existence of an SPE, it remains to verify whether the pair of inequalities (4) can be satisfied under some assumptions, even though the action sets of the players are non-rectangular in the max–min problem. Besides the issue of non-rectangular action sets, the main reason that the upper and lower values are different is mainly due to the ability of the minimizer to *deceive* the maximizer by altering the ranking of the most negative values. If we remove this ability from the network designer, we should expect that an SPE would exist. The following theorem makes this argument formal. Define $\gamma := \frac{4\|x_0\|_\infty^2}{\epsilon^2}$, $\epsilon > 0$. We assume that ϵ is chosen to guarantee $\gamma > 1$.

Theorem 6. *Given $\epsilon > 0$, assume that T is small enough such that (32) in the Appendix holds. Then, under Assumptions 1 and 2, a sufficient condition for the existence of an SPE for the underlying zero-sum game between the designer and the adversary is to select b such that*

$$0 \leq b \leq \min_{e_{ij}, e_{kl} \in \mathcal{E}} |\gamma a_{ij} - a_{kl}|, \quad (21)$$

given that $a_{ij} \neq a_{kl}$ and $a_{ij} > \gamma a_{kl}$ whenever $a_{ij} > a_{kl}$, for all $e_{ij}, e_{kl} \in \mathcal{E}$.

Proof: It suffices to show that $\mathcal{L}_\ell(v^*) = \mathcal{L}_\ell(0) = \mathcal{D}_\ell$ as this would imply that the adversary would break the same links whether he acts first or second, and as a result the strategy of the minimizer in both problems will be the same. This will guarantee that (4) is satisfied. This would occur if the minimizer cannot protect any of the links in $\mathcal{L}_\ell(0)$. In other words, this will happen if the minimizer cannot satisfy the condition in step 2 of Algorithm I for any $i \in \{1, \dots, \ell\}$. A sufficient condition for $\mathcal{L}_\ell(v^*) = \mathcal{L}_\ell(0) = \mathcal{D}_\ell$ to hold is to require

$$\min_{e_{ij} \in \Phi(\mathcal{P}(0))} (a_{ij} + b)\nu_{ij} > \max_{e_{ij} \in \mathcal{L}_\ell(0)} a_{ij}\nu_{ij}.$$

This implies that no matter how the designer changes the weights of the links in $\Phi(\mathcal{P}(0))$, he cannot make those links more negative than the links in $\mathcal{L}_\ell(0)$. To satisfy this inequality, we will establish that whenever $a_{ij}\nu_{ij} > a_{kl}\nu_{kl}$, we must have $(a_{ij} + b)\nu_{ij} > a_{kl}\nu_{kl}$, for all $e_{ij}, e_{kl} \in \mathcal{E}$. We can then re-write the condition on b as

$$b \leq \frac{a_{ij}\nu_{ij} - a_{kl}\nu_{kl}}{-\nu_{ij}} = a_{kl} \frac{|\nu_{kl}|}{|\nu_{ij}|} - a_{ij}, \quad \forall e_{ij}, e_{kl} \in \mathcal{E} \quad (22)$$

Consider the following two cases. If $\nu_{kl} \geq \nu_{ij}$, then we must have $a_{kl} > a_{ij}$. Then, by assumption we have that $a_{kl} > \gamma a_{ij}$. By Lemma 1 in the Appendix, we can write

$$a_{kl} \frac{|\nu_{kl}|}{|\nu_{ij}|} - a_{ij} \geq \frac{1}{\gamma} a_{kl} - a_{ij} > 0. \quad (23)$$

Next, consider the case when $\nu_{ij} > \nu_{kl}$. In this case, a_{ij} can be larger or smaller than a_{kl} . However, if $a_{ij} > a_{kl}$, and recalling that $a_{ij}\nu_{ij} > a_{kl}\nu_{kl}$, then

$$\gamma a_{kl} < a_{ij} < a_{kl} \frac{|\nu_{kl}|}{|\nu_{ij}|} \leq \gamma a_{kl},$$

which is a contradiction. The case $a_{kl} = a_{ij}$ is excluded by assumption. Hence, in this case, we must have $a_{ij} < a_{kl}$, and the inequality in (23) applies. Thus, by choosing b as in (21), we obtain the condition we are seeking. Note that we do not need to consider the case when $a_{ij}\nu_{ij} = a_{kl}\nu_{kl}$ since the players will be indifferent as to which link to choose. ■

Remark 4. *The condition derived in the above theorem requires the network to be “sufficiently diverse” in the sense that the weights of the links have to be not only different from each other, but also a factor γ apart. This is due to the fact that we were seeking uniform bounds on the ν_{ij} ’s, for all $e_{ij} \in \mathcal{E}$. If we allow b to vary with time, then one can find less restrictive conditions to ensure the existence of an SPE. However, this would require (22) to be verified at each time instant. Further, the bound derived in (23) is loose, because it was obtained by bounding $|v_{kl}|$ and $|v_{ij}|$ independently, for $e_{ij}, e_{kl} \in \mathcal{E}$. Tighter bounds could be given by studying the dynamics of $|\nu_{kl}|/|\nu_{ij}|$. However, studying the time derivative of this ratio is not tractable.*

Remark 5. *This result highlights the fact that, in general, Stackelberg games are more natural to study security problems than zero-sum games. In fact, the leader-follower formulation fits many real-world security scenarios; see [21] and the references therein. However, the sufficient condition we derive here is a step in the right direction for establishing the existence of an SPE for the zero-sum game between the designer and the adversary. We are currently investigating whether this condition is also necessary.*

V. CONCLUSION

In this paper, we have studied the impact of an adversarial attack on a network of agents performing consensus averaging. The adversary’s objective is to slow down the convergence of

the computation at the nodes to the global average. We introduced a network designer whose objective is to assist the nodes reach consensus by countering the attack of the adversary. The adversary and the network designer are capable of targeting links. We have formulated and solved two problems that capture the competition between the two players. We considered practical models for the players by constraining their actions along the problem horizon. The derived strategies were shown to exhibit a low worst-case complexity. When Zeno behavior is excluded, we showed that the optimal strategies admit a potential-theoretic analogy. Finally, we showed that when the link weights are sufficiently diverse, an SPE exists for the zero-sum game between the designer and the adversary.

Future work will focus on removing Assumption 1 and showing that Zeno behavior can be ruled out in optimality. Formulating the problems in discrete-time is also of interest. Another interesting line of research is to derive the optimal strategies when the knowledge of the players about the state and the topology of the network is restricted. When applying necessary conditions for optimality, e.g., the MP, to the min–max or the max–min problem, one must first prove the existence of optimal controllers. Such results can be viewed as existence results for equilibria in the general framework of Stackelberg games. This is another avenue for future research.

ACKNOWLEDGMENT

The authors would like to thank Prof. Behrouz Touri for valuable comments during the development of this work. We are also very grateful for the constructive comments made by the Associate Editor and the reviewers, which helped improve the original manuscript.

REFERENCES

- [1] A. Khamis, B. Touri, and T. Başar, “Robust distributed averaging on networks with adversarial intervention,” in *Proc. 52nd IEEE Conf. on Decision and Control (CDC)*, December 2013, pp. 7131–7136.
- [2] R. Olfati-Saber and R. M. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1520–1533, 2004.
- [3] R. Olfati-Saber, “Flocking for multi-agent dynamic systems: Algorithms and theory,” *IEEE Trans. Automat. Contr.*, vol. 51, no. 3, pp. 40–420, 2006.
- [4] A. Nedić, A. Ozdaglar, and A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Trans. Automat. Contr.*, vol. 55, no. 4, pp. 922–938, 2010.
- [5] M. O. Jackson and B. Golub, “Naive learning in social networks: Convergence, influence and wisdom of crowds,” *American Economic J.: Microeconomics*, vol. 2, no. 1, pp. 112–149, 2010.

- [6] A. Nedić and A. Ozdaglar, “Convergence rate for consensus with delays,” *J. Global Optimization*, vol. 47, no. 3, pp. 437–456, 2010.
- [7] L. Xiao, S. Boyd, and S.-J. Kim, “Distributed average consensus with least-mean-square deviation,” *J. Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.
- [8] B. Touri and A. Nedić, “Distributed consensus over network with noisy links,” in *Proc. 12th Internat. Information Fusion Conf.*, 2009, pp. 146–154.
- [9] A. Kashyap, T. Başar, and R. Srikant, “Quantized consensus,” *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.
- [10] S. Sundaram and C. Hadjicostis, “Distributed function calculation via linear iterative strategies in the presence of malicious agents,” *IEEE Trans. Automat. Contr.*, vol. 56, no. 7, pp. 1495–1508, July 2011.
- [11] A. Teixeira, H. Sandberg, and K. Johansson, “Networked control systems under cyber attacks with applications to power networks,” in *Proc. American Control Conference (ACC)*, July 2010, pp. 3690–3696.
- [12] A. Khanafer, B. Touri, and T. Başar, “Consensus in the presence of an adversary,” in *Proc. 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys)*, September 2012, pp. 276–281.
- [13] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
- [14] J. Rémy and C. Letamendia, *LTE Standards*, ser. ISTE. Wiley, 2014.
- [15] S. Goyal and A. Vigier, “Robust networks,” mimeo, University of Cambridge, Faculty of Economics, Tech. Rep., 2010.
- [16] H. J. LeBlanc, H. Zhang, S. Sundaram, and X. Koutsoukos, “Consensus of multi-agent networks in the presence of adversaries using only local information,” in *Proc. 1st Internat. Conf. High Confidence Networked Systems (HiCoNS)*, 2012, pp. 1–10.
- [17] F. Pasqualetti, A. Bicchi, and F. Bullo, “Consensus computation in unreliable networks: A system theoretic approach,” *IEEE Trans. Automat. Contr.*, vol. 57, no. 1, pp. 90–104, January 2012.
- [18] S. Kar and J. M. Moura, “Distributed consensus algorithms in sensor networks: Quantized data and random link failures,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, 2010.
- [19] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Philadelphia, U.S.: SIAM Series in Classics in Applied Mathematics, 1999.
- [20] J. Norris, *Markov Chains*. New York.: Cambridge Series in Statistical and Probabilistic Mathematics, 1997.
- [21] D. Korzhyk, Z. Yin, C. Kiekintveld, V. Conitzer, and M. Tambe, “Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness,” *Journal of Artificial Intelligence Research*, vol. 41, pp. 297–327, 2011.
- [22] R. Isaacs, *Differential Games*. New York: Wiley, 1965.

APPENDIX

A. Proof of Theorem 5

For a fixed strategy v of the designer, it was shown in [1] that the adversary’s strategy derived using the MP requires finding the lowest $f_{ij} = (a_{ij} + v_{ij})(p_i - p_j)(x_j - x_i)$ values, for all $e_{ij} \in \mathcal{E}$. However, Theorem 1 requires finding the lowest w_{ij} ’s. The designer’s strategy relies on finding the lowest $(p_i - p_j)(x_j - x_i)$ values according to the MP, and it requires finding the lowest ν_{ij} ’s according to Theorem 2. In order to prove the theorem, and since $w_{ij} = (a_{ij} + v_{ij})\nu_{ij}$,

$a_{ij} + v_{ij} \geq 0$, it is sufficient to show that $w_{ij} \leq w_{kl}$ implies that $f_{ij} \leq f_{kl}$, for all $e_{ij}, e_{kl} \in \mathcal{E}$. Without loss of generality, we will assume that $v_{ij} = v_{kl} = 0$. The Hamiltonian associated with the min–max problem is:

$$H(x, p, u, v) = \frac{1}{2}k(t)\|x(t) - \bar{x}\|_2^2 + p(t)^T A(u(t), v(t))x(t),$$

where p is the costate vector, whose existence is guaranteed by the MP because an optimal solution for the min–max exists. The first-order necessary conditions for optimality are (noting that $A^T = A$ and recalling that $V = V(0)$) [22]:

$$\begin{aligned} \dot{p} &= -\frac{\partial}{\partial x} H \\ &= -k(x - \bar{x}) - Ap, \quad p(T) = 0 \end{aligned} \tag{24}$$

$$\dot{x} = Ax, \quad x(0) = x_0 \tag{25}$$

$$u^*(v) = \arg \max_U H(x, p, u, v), \quad v^* = \arg \max_V H(x, p, u^*(v), v).$$

To prove the theorem, we will rely on approximating the state and costate up to first-order using Taylor expansion. To this end, we partition the problem's horizon into $L > K$ small sub-intervals of length $0 < \delta \leq \tau$, where τ was defined in (3), over which the system is time-invariant. More formally, define the times $0 = t_1 < t_2 < \dots < t_L < t_{L+1} = T$. Let A_i be the system matrix corresponding to the interval $[t_i, t_{i+1}]$, $i = 1, \dots, L$. We will denote the i -th row of matrix A_k by $A_{k,i}$ and its (i, j) -th element by a_{ij}^k . The proof comprises two steps: (i) we establish the claim of the theorem over $[t_L, t_{L+1}]$; and (ii) we generalize the argument to hold over $[0, T]$. We start by considering the interval $[t_L, t_{L+1}]$. The solutions to ODEs (24) and (25) over this interval are:

$$x_{A_L}(t) = e^{A_L(t-t_L)} x_{A_L}(t_L) \tag{26}$$

$$p_{A_L}(t) = \int_t^T e^{-A_L(t-\tau)} (x_{A_L}(\tau) - \bar{x}) d\tau. \tag{27}$$

Let $P_i(t) := e^{A_i t} = I + tA_i + \mathcal{O}(\delta^2)$. We can then re-write the above expressions as

$$\begin{aligned} x_{A_L}(t) &= P_L(t - t_L) x_{A_L}(t_L) \\ &= (I + (t - t_L)A) x_{A_L}(t_L) + \mathcal{O}(\delta^2) \\ p_{A_L}(t) &= \int_t^T P_L(\tau - t) [P_L(\tau - t_L) - M] x_{A_L}(t_L) d\tau \\ &= (T - t)(I - M) x_{A_L}(t_L) + \mathcal{O}(\delta^2), \end{aligned}$$

where the last equality follows because $(T-t)(T-t_L)A_L = \mathcal{O}(\delta^2)$. Define $\xi(\alpha, \beta) := \alpha - \beta$, $\alpha, \beta \in \mathbb{R}$, and write

$$x_{A_L}(t) = (I + \xi(t, t_L)A_L)x_{A_L}(t_L) + \mathcal{O}(\delta^2) \quad (28)$$

$$p_{A_L}(t) = \xi(T, t)(I - M)x_{A_L}(t_L) + \mathcal{O}(\delta^2). \quad (29)$$

Further, define the matrices $G := I + \xi(t, t_L)A_L$, $R := \xi(T, t)(I - M)$, and write

$$\begin{aligned} w_{ij} &= a_{ij}^L(x_{A_L, i} - x_{A_L, j})(x_{A_L, j} - x_{A_L, i}) \\ &= a_{ij}^L x_{A_L}(t_L)^T (G_i - G_j)(G_j - G_i)^T x_{A_L}(t_L) + \mathcal{O}(\delta^2) \\ f_{ij} &= a_{ij}^L x_{A_L}(t_L)^T (R_i - R_j)(G_j - G_i)^T x_{A_L}(t_L) + \mathcal{O}(\delta^2), \end{aligned}$$

where R_i^T, R_j^T are the i -th row of G and R , respectively. Using the above definitions, we obtain

$$\begin{aligned} (G_i - G_j)(G_j - G_i)^T &= -(I_i - I_j)(I_i - I_j)^T - \xi(t, t_L)((I_i - I_j)(A_{L, i} - A_{L, j})^T \\ &\quad + (A_{L, i} - A_{L, j})(I_i - I_j)^T) - \xi(t, t_L)^2(A_{L, i} - A_{L, j})(A_{L, i} - A_{L, j})^T. \end{aligned}$$

The last term is quadratic, and thus we can absorb it in $\mathcal{O}(\delta^2)$. We then have

$$\begin{aligned} a_{ij}^L(G_i - G_j)(G_j - G_i)^T - a_{kl}^L(G_k - G_l)(G_l - G_k)^T &= a_{kl}^L(I_k - I_l)(I_k - I_l)^T \\ - a_{ij}^L(I_i - I_j)(I_i - I_j)^T + (a_{kl}^L(I_k - I_l)(A_{L, k} - A_{L, l})^T - a_{ij}^L(I_i - I_j)(A_{L, i} - A_{L, j})^T)\xi(t, t_L) \\ + (a_{kl}^L(A_{L, k} - A_{L, l})(I_k - I_l)^T - a_{ij}^L(A_{L, i} - A_{L, j})(I_i - I_j)^T)\xi(t, t_L) + \mathcal{O}(\delta^2). \end{aligned}$$

Similarly, we have

$$\begin{aligned} a_{ij}^L(R_i - R_j)(G_j - G_i)^T - a_{kl}^L(R_k - R_l)(G_l - G_k)^T &= (a_{kl}^L(I_k - I_l)(I_k - I_l)^T \\ - a_{ij}^L(I_i - I_j)(I_i - I_j)^T)\xi(T, t) + \mathcal{O}(\delta^2). \end{aligned}$$

Let $\Gamma_1 = a_{kl}^L(I_k - I_l)(I_k - I_l)^T - a_{ij}^L(I_i - I_j)(I_i - I_j)^T$ and $\Gamma_2 = a_{kl}^L(I_k - I_l)(A_{L, k} - A_{L, l})^T - a_{ij}^L(I_i - I_j)(A_{L, i} - A_{L, j})^T$. We now have

$$\begin{aligned} w_{ij} - w_{kl} &= x_{A_L}(t_L)^T (\Gamma_1 + \xi(t, t_L)\Gamma_2 + \xi(t, t_L)\Gamma_2^T)x_{A_L}(t_L) + \mathcal{O}(\delta^2), \\ f_{ij} - f_{kl} &= \xi(T, t)x_{A_L}(t_L)^T \Gamma_1 x_{A_L}(t_L) + \mathcal{O}(\delta^2). \end{aligned}$$

If $w_{ij} - w_{kl} \leq 0$, since $\xi(T, t) \geq 0$, we can write

$$\xi(T, t)(w_{ij} - w_{kl}) = x_{A_L}(t_L)^T (\xi(T, t)\Gamma_1 + \xi(T, t)\xi(t, t_L)\Gamma_2 + \xi(T, t)\xi(t, t_L)\Gamma_2^T)x_{A_L}(t_L) + \mathcal{O}(\delta^2) \leq 0,$$

or $\xi(T, t)x_{A_L}(t_L)^T \Gamma_1 x_{A_L}(t_L) + \mathcal{O}(\delta^2) \leq 0$, but the left hand side is $f_{ij} - f_{kl}$; hence, $w_{ij} \leq w_{kl} \implies f_{ij} \leq f_{kl}$ as required. So far, we have verified the claim of the theorem over the interval $[t_L, T]$ only. We are now in a position to generalize the statement of the theorem to the interval $[0, T]$. The only complication that arises when studying this interval is that the terminal condition, i.e. $p_{L-1}(t_L)$, is not forced to be zero as in $[t_L, T]$. Over the interval $[t_{L-1}, t_L]$, the state and costate are

$$\begin{aligned} x_{L-1}(t) &= e^{A_{L-1}(t-t_{L-1})}x_{A_{L-1}}(t_{L-1}) \\ p_{A_{L-1}}(t) &= e^{-A_{L-1}(t-t_{L-1})}p_{A_{L-1}}(t_{L-1}) - \int_{t_{L-1}}^t e^{-A_{L-1}(t-\tau)}(x_{A_{L-1}}(\tau) - \bar{x})d\tau. \end{aligned}$$

Solving for $p_{A_{L-1}}(t_{L-1})$ in terms of $p_{A_{L-1}}(t_L)$ and substituting back, we can write $p_{A_{L-1}}(t)$ in terms of $p_{A_{L-1}}(t_L)$ as follows:

$$p_{A_{L-1}}(t) = e^{-A_{L-1}(t-t_L)}p_{A_{L-1}}(t_L) + \int_t^{t_L} e^{-A_{L-1}(t-\tau)}(x_{A_{L-1}}(\tau) - \bar{x})d\tau. \quad (30)$$

By continuity of the state and costate functions, it follows that $x_{A_{L-1}}(t_L) = x_{A_L}(t_L)$, $p_{A_{L-1}}(t_L) = p_{A_L}(t_L)$. Using a first-order Taylor expansion and (29), we can write

$$\begin{aligned} p_{A_{L-1}}(t) &= (I + \xi(t_L, t)A_{L-1})p_{A_L}(t_L) + \xi(t_L, t)(I - M)x_{A_{L-1}}(t_{L-1}) + \mathcal{O}(\delta^2) \\ &= \xi(t_L, t)(I - M)x_{A_{L-1}}(t_L) + \xi(t_L, t)(I - M)x_{A_{L-1}}(t_{L-1}) + \mathcal{O}(\delta^2). \end{aligned}$$

We can further simplify this expression using $x_{A_{L-1}}(t)$ as follows:

$$\begin{aligned} \xi(t_L, t)(I - M)x_{A_{L-1}}(t_L) &= \xi(t_L, t)(I - M)e^{A_{L-1}(t_L-t_{L-1})}x_{A_{L-1}}(t_{L-1}) \\ &= \xi(t_L, t)(I - M)x_{A_{L-1}}(t_{L-1}) + \mathcal{O}(\delta^2), \end{aligned}$$

and therefore we have

$$p_{A_{L-1}}(t) = 2\xi(t_L, t)(I - M)x_{A_{L-1}}(t_{L-1}) + \mathcal{O}(\delta^2). \quad (31)$$

Comparing (29) and (31), we conclude that the argument used to prove the claim over the interval $[t_L, T]$ applies over $[t_{L-1}, t_L]$. Hence, $w_{ij} - w_{kl} \leq 0$ implies that $f_{ij} - f_{kl} \leq 0$ over $[t_{L-1}, t_L]$.

Note that we can generalize (30) to any interval $[t_i, t_{i+1}]$, $i = 1, \dots, L$, as follows:

$$p_{A_i}(t) = e^{-A_i(t-t_{i+1})}p_{A_i}(t_{i+1}) + \int_t^{t_{i+1}} e^{-A_i(t-\tau)}(x_{A_i}(\tau) - \bar{x})d\tau.$$

Following similar steps to the above, we can arrive at

$$p_{A_i}(t) = \frac{T - t_i}{\delta} \xi(t_{i+1}, t)(I - M)x_{A_i}(t_i) + \mathcal{O}(\delta^2), \quad t \in [t_i, t_{i+1}],$$

which maintains the same structure as in (31), and the claim therefore holds for the interval $[t_i, t_{i+1}]$, $i = 1, \dots, L$, and the theorem is proved.

B. Technical Results

Proposition 1. *Given τ_1, τ_2, τ_3 , which were defined in terms of $\delta > 0$ in Theorem 1, let f be a real-valued function. Then, if $f(\delta) = \mathcal{O}(\tau_i^2)$ as $\delta \rightarrow 0$, we have $f(\delta) = \mathcal{O}(\delta^2)$, $i \in \{1, 2, 3\}$. Also, if $f(\delta) = \tau_i \mathcal{O}(\tau_j^2)$ as $\delta \rightarrow 0$, then $f(\delta) = \mathcal{O}(\delta^3)$, $i, j \in \{1, 2, 3\}$.*

Proof: Recall that we write $f(x) = \mathcal{O}(g(x))$, for some real-valued function g , as $x \rightarrow a$ if there exist constants M, γ such that $|f(x)| \leq M|g(x)|$, for all x satisfying $|x - a| < \gamma$. Since $f(\delta) = \mathcal{O}(\tau_i^2)$ as $\delta \rightarrow 0$, and recalling that by definition we have $\tau_i \leq \delta$ for $i \in \{1, 2, 3\}$, we can write $f(\delta) \leq M\tau_i^2 \leq M\delta^2$. Hence, $f(\delta) = \mathcal{O}(\delta^2)$. To prove the second statement, recall that $h(x)\mathcal{O}(g(x)) = \mathcal{O}(h(x)g(x))$, for any two real-valued functions h, g . Hence, as $\delta \rightarrow 0$, we have $f(\delta) = \tau_i \mathcal{O}(\tau_j^2) = \mathcal{O}(\tau_i \tau_j^2)$. Therefore, $f(\delta) \leq M\tau_i \tau_j^2 \leq M\delta^3$ and $f(\delta) = \mathcal{O}(\delta^3)$. ■

Lemma 1. *Given $\epsilon > 0$ and $\delta \leq \tau$, τ defined in (3), one can select the problem horizon T small enough such that*

$$\epsilon \leq |x_i(t) - x_j(t)| \leq 2\|x_0\|_\infty, \quad \forall e_{ij} \in \mathcal{E}, \quad (32)$$

for all $t \in [0, T]$.

Proof: By the structure of the system matrix in (1), we can deduce that $|x_i - x_j|$ cannot increase as $t \rightarrow T$. Thus

$$\begin{aligned} |x_i(t) - x_j(t)| &\leq \max_{1 \leq i, j \leq n} |x_i(0) - x_j(0)| \\ &\leq 2 \max_{1 \leq i \leq n} |x_i(0)| = 2\|x_0\|_\infty. \end{aligned}$$

This provides the uniform upper bound. In order to obtain a uniform lower bound, we need to ensure that $|x_i(t) - x_j(t)|$ does not approach zero as $t \rightarrow T$. We are seeking a time t^* such that for a given $\epsilon > 0$, we have $|x_i(t) - x_j(t)| \geq \epsilon$ for all $t < t^*$ and all $e_{ij} \in \mathcal{E}$. We can then fix $T < t^*$ to ensure the existence of a uniform lower bound on $|x_i(t) - x_j(t)|$. Let us again restrict our attention to a small interval $[t_0, t_0 + \delta]$ where the system is time-invariant, and let the system matrix over this interval be A . We require that the system did not reach equilibrium over this

interval, i.e., $x(t_0 + \delta) \neq \bar{x}$. Without loss of generality, we assume that $x_1(t_0) > \dots > x_n(t_0)^2$.

Define the following dynamics

$$\frac{d}{dt}(\bar{y}_i - x_1(t_0)) = \sum_{j \neq i} A_{ij}(x_1(t_0) - \bar{y}_i), \quad \frac{d}{dt}(\underline{y}_i - x_n(t_0)) = \sum_{j \neq i} A_{ij}(x_n(t_0) - \underline{y}_i),$$

with initial conditions $\bar{y}_i(t_0) = 2x_1(t_0)$, $\underline{y}_i(t_0) = 2x_n(t_0)$. Note that $\dot{x}_i = \sum_{j \neq i} A_{ij}(x_j - x_i)$. It follows that $\dot{\underline{y}}_i \leq \dot{x}_i \leq \dot{\bar{y}}_i$. By the comparison principle, we conclude that $\underline{y}_i - x_n(t_0) \leq x_i \leq \bar{y}_i - x_1(t_0)$, for $i \in \mathcal{N}$. Note that we can readily find the solution trajectories for \bar{y} and \underline{y} . By defining $a_i = \sum_{j \neq i} A_{ij}$, we can then write

$$\bar{y}_i - x_1(t_0) = e^{-a_i(t-t_0)}x_1(t_0), \quad \underline{y}_i - x_n(t_0) = e^{-a_i(t-t_0)}x_n(t_0).$$

By solving the equation $\bar{y}_{i-1} - x_1(t_0) = \underline{y}_i - x_n(t_0)$, we can find a time t_i^* when x_{i-1} can potentially meet x_i :

$$t_i^* = \frac{1}{a_{i-1} - a_i} \ln \left(\frac{x_1(t_0)}{x_n(t_0)} \right) + t_0.$$

If $t_i^* > t_0 + \delta$, for all $i \in \mathcal{N}$, then we need to propagate the solution forward, and keeping in mind that the system matrix could change, until we find a time t_i^* in some interval $[\tilde{t}, \tilde{t} + \delta]$ where $\bar{y}_{i-1} = \underline{y}_i$ for some $i \in \mathcal{N}$. Then, for a given $\epsilon > 0$, we can select $T < t_i^*$ such that $|x_i - x_{i-1}| \geq |\underline{y}_i - \bar{y}_{i-1}| \geq \epsilon$; hence, we conclude that for this choice of T we can guarantee that $|x_i - x_j| \geq \epsilon > 0$ for all $e_{ij} \in \mathcal{E}$. ■

²We are making the implicit assumption that $x_1(0) > \dots > x_n(0)$.